SCIENCE CHINA
Life Sciences

• CUSBEA ARTICLE SERIES •
• RESEARCH PAPER •

August 2012 Vol.55 No.8: 709–725
doi: 10.1007/s11427-012-4350-7

Primate phylogeny: molecular evidence for a pongid clade excluding humans and a prosimian clade containing tarsiers

HUANG Shi

State Key Laboratory of Medical Genetics, Central South University, Changsha 410078, China

Received June 10, 2012; accepted June 28, 2012

Unbiased readings of fossils are well known to contradict some of the popular molecular groupings among primates, particularly with regard to great apes and tarsiers. The molecular methodologies today are however flawed as they are based on a mistaken theoretical interpretation of the genetic equidistance phenomenon that originally started the field. An improved molecular method the 'slow clock' was here developed based on the Maximum Genetic Diversity hypothesis, a more complete account of the unified changes in genotypes and phenotypes. The method makes use of only slow evolving sequences and requires no uncertain assumptions or mathematical corrections and hence is able to give definitive results. The findings indicate that humans are genetically more distant to orangutans than African apes are and separated from the pongid clade ~17.6 million years ago. Also, tarsiers are genetically closer to lorises than simian primates are. Finally, the fossil times for the radiation of mammals at the K/T boundary and for the Eutheria-Metatheria split in the Early Cretaceous were independently confirmed from molecular dating calibrated using the fossil split times of gorilla-orangutan, mouse-rat, and opossum-kangaroo. Therefore, the re-established primate phylogeny indicates a remarkable unity between molecules and fossils.

genetic non-equidistance, genetic equidistance, molecular clock, Neutral theory, MGD hypothesis, slow clock, pongid, tarsiers, orangutans, chimpanzees, gorillas

Citation: Huang S. Primate phylogeny: molecular evidence for a pongid clade excluding humans and a prosimian clade containing tarsiers. Sci China Life Sci, 2012, 55: 709−725, doi: 10.1007/s11427-012-4350-7

When fossils were interpreted without any biased influence from molecular data, the consensus view was that human is the outgroup to a pongid (orangutan-gorilla-chimpanzees) clade and diverged from pongids ~18 million years (Myr) ago [1−6]. In contrast, popular interpretations of molecular data suggest that humans and chimpanzees belong to the same clade to the exclusion of other great apes and shared a common ancestor merely 5 Myr ago [7–9].

The genetic equidistance result of Margoliash, which—together with those of Zuckerkandl and Pauling in 1962—directly inspired the molecular clock hypothesis and in turn the Kimura Neutral theory of macroevolution, shows that different species are approximately equidistant to a simpler outgroup in protein sequence similarity [10]. Recent work shows that this equidistance result has another characteristic, the overlap feature, which invalidates the clock/Neutral theory interpretation [11]. A position where two or more species have each had a substitution event is termed an overlap (or saturated or coincident substitution) position, unlike simple substitution, in which only one species has changed (Figure 1A, species A and B have 6 overlap positions). The genetic equidistance phenomenon minimally requires three species for sequence alignment, including two sister species and a less derived outgroup. The overlap feature shows several overlapped coincidental substitutions where any pair of these three species is different in sequence. If after speciation, two species randomly accumulate substitutions with similar rate as assumed by the clock/Neutral theory, then the chance for a substitution in one species to occur coincidentally at the same overlap po-

email: huangshi@sklmg.edu.cn

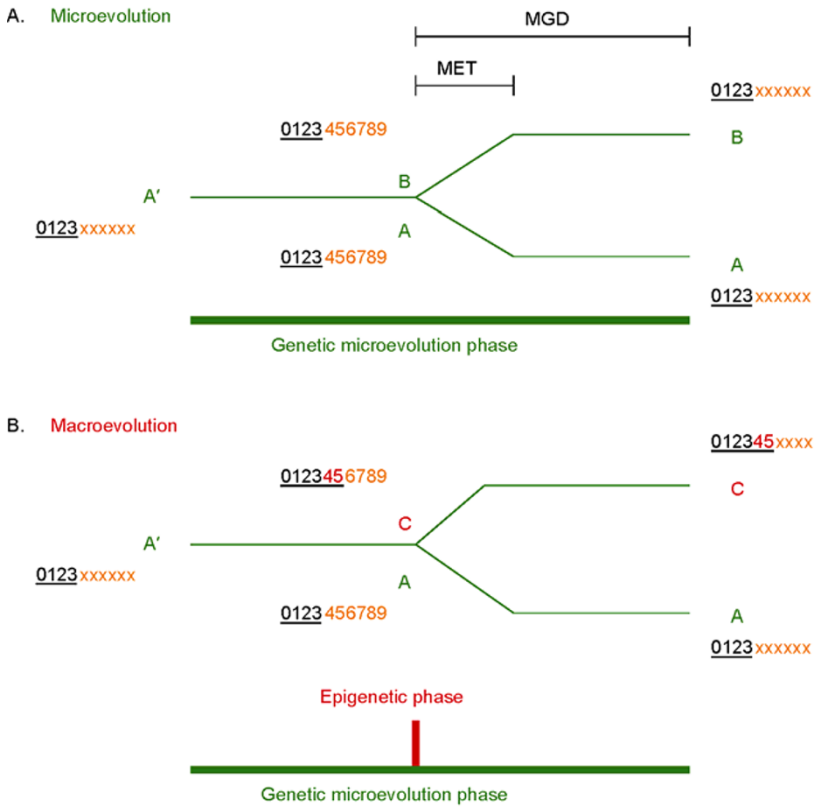


Figure 1 Evolutionary events according to the Maximum Genetic Diversity hypothesis. A 10 amino acid peptide with each position represented by a number is used to illustrate the mutation events during evolution. X represents any amino acid. A mismatch is indicated by X vs. X or X vs. *N* (*N*=0, 1, 2, etc.). Underlined positions represent non-changeable residues due to functional or epigenetic restriction. A, Microevolution. The ancestor species A′ produced a pair of sister individuals A and B, which shared the same sequence for an orthologous gene. The genetic distance between A and B gradually increases until reaching a plateau or MGD, as indicated by the branching lines that flatten near the end as well as by the large number of overlapped coincidental substitutions (6 such overlaps are shown). Relative to the MGD, the MET is an incomplete theory that cannot cover the time period after the plateau. B, Macroevolution. The ancestor species A′ undergoes microevolution and gradually reaches some level of genetic diversity, during which time nearly every genome variation allowed within the MGD may have a chance to exist for a while. When one of these variations also happens to be compatible with a higher level of epigenetic complexity such as the genome as shown for sister individuals A and C at the beginning of speciation, a punctuational increase in epigenetic complexity would take place in one of these sisters such as C, which in turn reduced the number of changeable positions from 6 to 4. After the epigenetic phase at the beginning of speciation, the genetic microevolutionary phase immediately follows that would gradually create greater genetic distance between A and C until reaching a plateau distance of 60% nonidentity which is determined by the mutations in the less derived species A.

sition where the other species also has a substitution should largely follow probability theory. Indeed, for microevolution of similar species such as among different strains of yeasts, the number of overlap positions relative to the total number of substitutions is small and consistent with probability calculation. In contrast, for macroevolution of distinct species of different biological complexity such as yeast versus drosophila, the number of overlap positions is much greater than expected by chance. Thus, the overlap feature shows a clear distinction between macroevolution and microevolution, where macroevolution is mostly about major changes in organismal complexity whereas microevolution is not.

The Neutral theory trivializes Darwinian natural selection and disconnects genotypes and phenotypes. It was originally a population genetics theory and turned into a macroevolution theory by Kimura when he used it to explain the molecular clock, which assumes macroevolution to be the same as population genetics and microevolution [12]. The Modern Evolution Theory (MET) combines Darwin's and Kimura's theory. The Maximum Genetic Diversity (MGD) hypothesis is an alternative model that more tightly unites genotypes and phenotypes by greatly reducing the extent of neutral sequences [13,14]. Genetic diversity, defined here as percent position difference in the aligned sequence of homologous proteins or DNA created by point mutations, has a loose, inverse correlation with epigenetic complexity, defined as the total number of cell types and epigenetic molecules. Genetic diversity cannot increase indefinitely with time and has a maximum limit being restricted by function or epigenetic complexity. The MGD of simple organisms is greater than that of complex organisms.

The MGD hypothesis defines microevolution and macroevolution differently from the standard MET definition and considers them different (Figure 1A vs. 1B). Macroevolution involves major changes in epigenetic complexity but

microevolution does not. Macroevolution involves a fast and punctuational epigenetic event whereas microevolution is largely a slow process of random point mutations and similar to population genetics. Macroevolution automatically includes microevolutionary mechanisms as part of the speciation process since the events following the epigenetic change are largely microevolutionary (Figure 1B). The MGD follows the MET for microevolution as well as for the microevolutionary phase of macroevolution and use it to cover evolutionary processes at linear phases where genetic distance/diversity has yet to reach maximum limit (Figure 1). For the plateau phases, the MGD with MET as a part of it provides a more accurate account. For the epigenetic phase of macroevolution, however, distinct from MET, the MGD posits that the genetic diversity of the more complex species would be reduced as a result of increase in epigenetic complexity (Figure 1B).

The overlap feature of the genetic equidistance result of Margoliash is evidence of fast evolving genes reaching MGD [11]. For macroevolution over long time scales when fast evolving genes have reached MGD, sequence difference in these genes between two species of different complexity is a reflection of the MGD of the simple species (Figure 2A). In contrast, for microevolution of short time scale or for slow evolving genes where the molecular clock holds and distance is still linearly related to time, the number of overlap positions would be small and consistent with chance. We can designate the former "maximum genetic equidistance" and the latter "linear genetic equidistance", which can be easily distinguished by the overlap ratio (Figure 2B). The overlap ratio is defined as the number of actual overlap positions divided by the number of candidate positions that in any three species comparison involving an outgroup include all the different positions between two sister lineages [11].

Thus, for fast evolving genes, the MGD predicts either genetic equidistance or non-equidistance to an outgroup depending on the epigenetic complexity of the outgroup (Figure 3, predictions 1 and 2). The genetic distance in fast evolving genes between a complex outgroup and a simple taxon is mainly determined by the mutations and MGD of the simple taxon. If one of the sister taxa is more complex than the others, it would have lower MGD and would show higher sequence similarity to a more complex outgroup, resulting in the phenomenon of molecular/genetic non-equidistance of the sister species to the more complex outgroup in fast evolving genes. Here, I present novel evidence for this phenomenon or prediction 2. Evidence for prediction 1 is the genetic equidistance phenomenon, which has been well documented [10,15,16].

The MGD hypothesis predicts the phenomenon of genetic non-equidistance to a taxon only in slow but not in fast evolving sequences given non-equidistance in time to this taxon of two other species that are not less complex than the taxon (Figure 3, predictions 3 and 4; Figure 4). Here, slow evolving genes display good correlation between distance and time but fast evolving genes do not as they have undergone substitutions to the maximum possible. Evidence for this novel phenomenon or predictions 3 and 4 in Figure 3 is here shown.

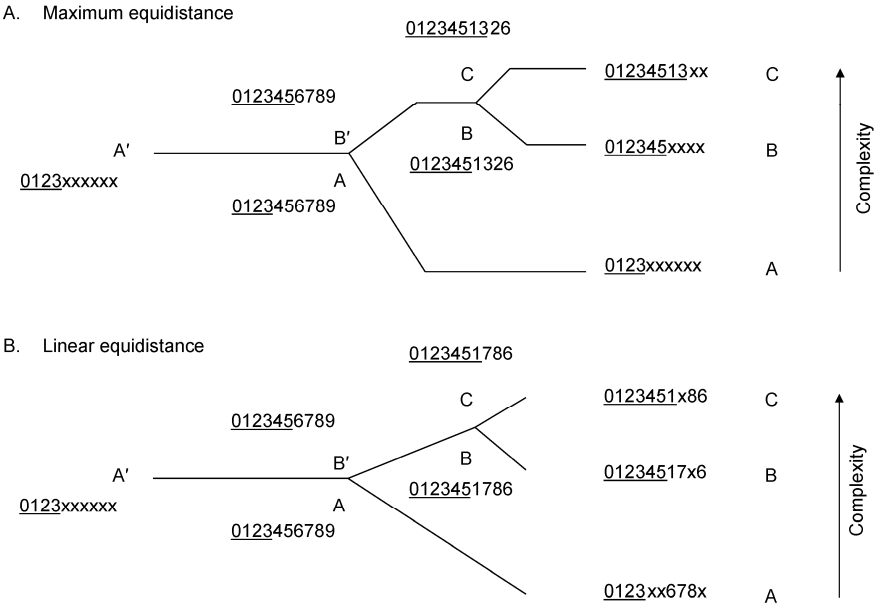
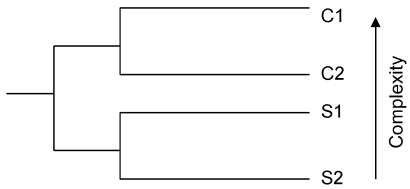


Figure 2 The genetic equidistance result and the MGD hypothesis. A, Maximum genetic equidistance. Similar to Figure 1, a 10 amino acid peptide is used to illustrate the evolution process. When the protein is fast evolving, the observed equidistance today would be maximum distance with a large overlap ratio. The figure shows 4 overlap positions with an overlap ratio 1. The distance of C-A is 60%, the same as that of B-A. This is a schematic representation of the original Margoliash genetic equidistance result. B, Linear genetic equidistance. When the protein is slow evolving, assuming molecular clock holds, the observed equidistance today would be linear distance with a small overlap ratio. Here every substitution in any species would mean an increase in distance. The figure shows 0 overlap position with an overlap ratio 0. The distance of C-A is 50% and equals that of B-A.



Predictions by the clock/neutral theory:

1. Equidistance to a simpler outgroup
   Dist. C1-S2 = Dist. C2-S2
2. Equidistance to a complex outgroup
   Dist. S2-C1 = Dist. S1-C1
3. Non-equidistance to a simpler taxon in slow evolving genes
   Dist. C1-S2 > Dist. S1-S2
4. Same as 3 for fast evolving genes.

Predictions by the MGD hypothesis:

1. Same as 1 above for fast evolving genes
2. Non-equidistance to a complex outgroup in fast evolving genes
   Dist. S2-C1 > Dist. S1-C1
3. Same as 3 above
4. Equidistance to a simpler taxon in fast evolving genes
   Dist. C1-S2 = Dist. S1-S2

Figure 3 Genetic distance to an outgroup as predicted by the molecular clock/neutral theory and the MGD hypothesis. Two clades of organisms are shown with the top one more complex, and two taxa within each clade are shown with the organism on top more complex. Both the molecular clock/neutral theory and the MGD hypothesis can make 4 predictions as shown. The two hypotheses differ only in predictions 2 and 4.

Traditional molecular phylogeny methods are based on the Neutral theory that is in turn based on the molecular clock. Therefore, existing molecular phylogeny methods are either explicitly or implicitly based on the flawed molecular clock concept, which have created some major contradictions with paleontological results, including, just for the mammals, the position of great apes and tarsiers, the timing of mammal radiation, and the split between Eutheria and Metatheria mammals [17–26]. It is also easily notable that traditional molecular methods have all self-proven themselves incorrect by repeatedly turning solid factual data into conflicting interpretations of reality, one of which must be false. One good example of endless conflicting results is the position of tarsiers [27]. A correct method should either produce only correct results or no results if informative data are not available. The Neutral theory would be fine for inferring phylogeny if only slow evolving neutral sequences are used. Here, I developed the slow clock method to re-establish a correct primate phylogeny.

1 Materials and methods

1.1 Sequence selection and alignments

Protein sequences from a specific taxon were retrieved from the NCBI protein database. Homology comparisons were performed using BLASTP on the NCBI server. Only orthologs were selected for comparison based on high sequence identity and reciprocal BLAST.

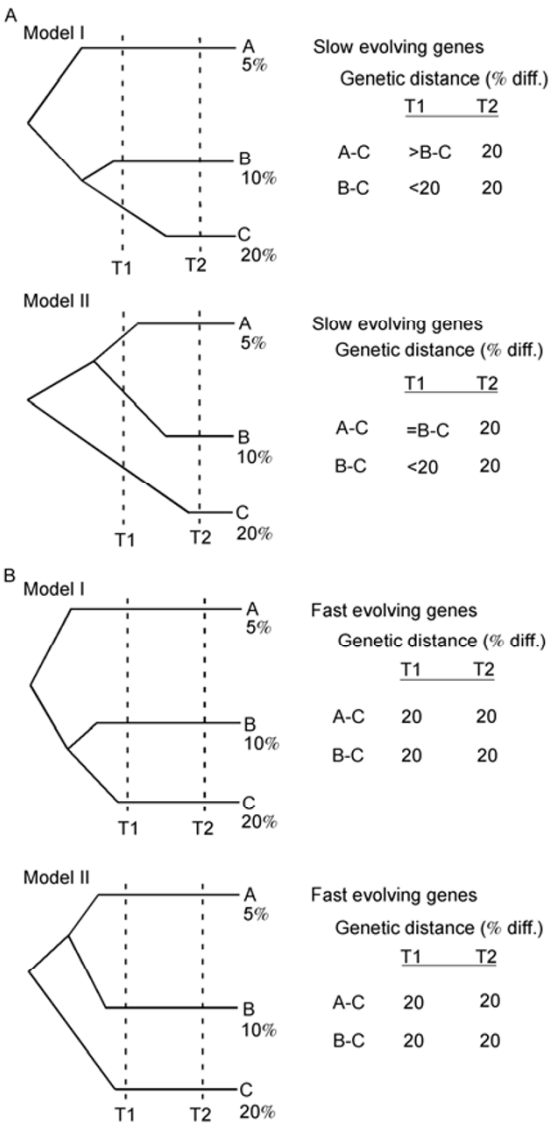


Figure 4 Inferring genealogy from sequence similarity in slow evolving neutral sequences. For any given three species A, B, and C of different epigenetic complexity, with A least complex and having low MGD (say, 5% sequence dissimilarity for a given protein) and B higher (10%) and C still higher (20%), there are two possible phylogenetic models as shown. A, Slow evolving neutral sequences can distinguish the two models. The two models predict different results for slow evolving neutral sequences at a time (T1) when the genetic distances have not yet reached the maximum. B, Fast evolving sequences cannot distinguish the two models.

1.2 Genetic equidistance test

Each gene was randomly selected from the NCBI database without any intentional bias or intent to influence in any biased way the outcome of the equidistance test. If A and B are equidistant (or non-equidistant) to C at the whole genome level, then a random sampling of a small set of the

genome should show the same. Equidistance means that, while some genes may show exact equidistance, some would show approximate equidistance. For genes that show approximate equidistance, the number of genes with greater similarity between A and C than between B and C should be similar to the number of genes with less similarity between A and C than between B and C (P>0.05).

The method relies on the availability of a set of randomly selected genes that is large enough for reaching statistical significance. While the availability of a gene sequence in the GenBank has specific reasons and hence is not strictly random, none of the reasons is in anyway linked to the equidistance test. Any non-biased selection scheme of these genes would satisfy the randomness requirement of the equidistance testing method here. The enrollment of genes for a test was stopped when the number of genes already enrolled was enough for drawing statistically significant conclusions. No gene was either included in or excluded from a test after knowing its effect on the test result.

The classification of a gene as fast or slow evolving was made after the enrollment of the gene for any given test. The cutoff score in percentage identity was arbitrarily made for each test so that the number of fast evolving genes is approximately similar to that of slow evolving genes to ensure that each set has sufficient number of genes for statistical testing. For inferring genealogy by equidistance testing to a simpler taxon, an internal control for randomness of gene selection is that the set of fast evolving genes should give equidistance result. For the equidistance test, non-informative genes include those that have no orthologous GenBank sequences in one of the concerned taxa, have long alignment gaps, are identical among the taxa, show exact equidistance from the outgroup, under strong positive selection (for example, major histocompatibility complex genes), or have many polymorphisms that prevent meaningful inference of equidistance.

1.3 Overlap ratio test

Additionally, slow genes should show less number of overlapped or coincident substitutions (Figure 2B). To calculate overlap ratio of a gene as defined by Huang [11], two sister species and an outgroup were aligned. The absence of any amino acid positions where any pair of the three species is different indicates an overlap ratio 0. Genes with overlap ratio >0 were non-informative and excluded from the calculation of divergence times.

1.4 Calculation of divergence time

Calculation of human-orangutan divergence time based on the gorilla fossil split time of 12 Myr ago was performed using the formula: Divergence time of human and orangutan=12×the distance for any given protein between human and orangutan divided by the distance between gorilla and orangutan. The divergence time for other species were calculated similarly. Only slow genes with overlap ratio=0 were used for calculation, where the distance is linearly related to time and need no uncertain mathematical corrections since every substitution event means an increase of 1 unit of distance (Figure 2B).

1.5 Statistical methods

Statistical methods used were Student's t test and Fisher's exact test, 2 tailed.

2 Results

2.1 Genetic non-equidistance to a more complex outgroup despite equidistance in time

To test prediction 2 in Figure 3, human was used as the outgroup to compare with sister species from a simpler group. For each group, where possible, two sister species were identified with one representing a simple organism and the other more complex. Genetic equidistance of A and B to an outgroup C can be established if the number of genes showing greater similarity between A and C than between B and C is similar to the number of genes showing less similarity between A and C than between B and C (P>0.05). For many of the analyses here, mitochondrial proteins were used because they are the only sequences available.

2.1.1 The mollusk phylum

The bivalves have existed since the Cambrian period. The octopuses have complex nervous systems and are considered among the most intelligent invertebrates. As shown in Table S1, a sampling of 10 mitochondrial proteins showed that humans are significantly closer to octopus (*Octopus vulgaris*) than to cockle (*Acanthocardia tuberculatum*) (10 showed more similarity between human and octopus than between human and cockle while 0 showed less, P<0.05).

2.1.2 The brachiopod phylum

The inarticulate brachiopod genus Lingula (*Lingula anatina*) is the oldest, relatively evolutionarily unchanged animal known since 550 Myr ago. Terebratulids (*Terebratulina retusa*) are modern articulate brachiopods and appeared 430 Myr ago. As shown in Table S2, humans are closer to Terebratulina than to Lingula (P<0.05). Lingula is equidistant to Terebratulina and human (P=0.64).

2.1.3 The reptile group (including birds)

Snakes maybe simple reptiles without limbs whereas birds have complex flying capacities. A sampling of 10 mitochondrial proteins shows that snakes are significantly more distant to humans than birds are (P<0.05). A random sampling of 13 nuclear genes also showed the same result

(P<0.05) (Table S3).

2.1.4 Other major groups of organisms

As shown in the Tables S4–S11, significant non-equidistance to humans was found for sister species within the teleost fish clade, the arthropod phylum, the Porifera phylum, and the fungi kingdom, but was not found for the amphibian group, the Echinoderm phylum, the Annelida phylum, the Nematode phylum, the Platyhelminthes phylum, the Cnidaria phylum, the Plant kingdom, the protist alveolates superphylum, and the bacteria kingdom. The failure to detect non-equidistance could be due to several reasons, including insufficient number of sampled genes or species, little difference in epigenetic complexity among sister species, and emergence of group-specific domains since separating from humans but before divergence of sister species within the group such as plants.

In all five cases (except plants) where difference in complexity of the sister species can be inferred (octopus vs. cockle, Terebratulina vs. Lingula, bird vs. snake, dragonfly vs. louse, and smut vs. yeast), the more complex species always show greater sequence similarity to humans in fast evolving genes, fully conforming to the predictions of the MGD hypothesis but not that of the clock/Neutral theory (Figure 3, prediction 2).

2.2 Difference between slow and fast evolving sequences in phylogeny inference

The phenomenon of genetic non-equidistance as shown above indicates that, for any three species, A, B, and C, with A being most complex and C least complex, a smaller distance between A and B relative to A and C cannot be used to group A and B to the exclusion of C. To infer genealogy, one must rely on the genetic distance to C as measured by slow evolving genes that contain slow evolving neutral sequences (Figure 4A, time T1). Only when A and B are equidistant to C in slow evolving neutral sequences, they can be grouped in the same clade to the exclusion of C (Figure 4A, Model II). If, however, B is closer to C than A is, then B and C would belong to the same clade to the exclusion of A (Figure 4A, Model I).

Slow evolving genes are those that show high identity between the simpler taxon C and a more complex taxon that is most similar to C in phenotypes. If B is more similar to C than A is, then B should be used for comparison with C to identify slow evolving genes. Large dissimilarity in phenotypes between A and C may indicate longer time of separation. Thus, relative to a list of high identity genes between B and C, a list of high identity genes between A and C would contain more genes that have reached MGD.

The genetic distance of A or B to C in slow evolving genes is mainly determined by the neutral mutation rate of C within the neutral diversity range of C (i.e., 20% for the example in Figure 4). Since the neutral mutation rate of C should be roughly constant over evolutionary time, the genetic distance of A or B to C should reflect the time of separation with C. In Model I of Figure 4A, knowing the mutation rate of C based on the fossil split time of B (or A) can be used to calculate the split time for A (or B). Here fast evolving genes should not be used as they would have reached MGD and would show that C is equidistant to A and B even if B and C belong to the same clade (Figure 4B).

In addition to high identity, slow evolving genes must show smaller number of overlap positions or coincident substitutions so that almost every substitution contributes to the distance. They must also be different from low identity genes in an equidistance test where the simpler taxon C is compared to a sister species and a complex outgroup. Most histone lysine methyltransferases (KMTs) (6 of 9) have identities between zebrafish and pufferfish that are equal to or slightly lower than that between zebrafish and human or mouse, showing that these proteins have reached the MGD for fishes (Table S12). In contrast, only 2 of 12 ribosomal proteins have reached the MGD. Thus, the KMT family is significantly different from the ribosome family in having more proteins reaching MGD (P=0.03). This correlates well with the fact that the average identity between the two fishes for the KMT family is significantly smaller than that of the ribosome family (65.1±8.5 vs. 92.1±4.7, P<0.001). Here, only slow evolving genes are informative: human is the outgroup because zebrafish is closer to pufferfish than to human in slow evolving genes such as ribosomal proteins. This result confirms predictions 3 and 4 by the MGD hypothesis (Figure 3).

2.3 Primate phylogeny

2.3.1 Humans are the sister taxon to a pongid clade

To use slow evolving neutral sequences to measure distance to the least derived or least complex species is here termed the "slow clock" method of phylogeny inference. I randomly picked a set of orangutan proteins to determine whether gorillas or chimpanzees are closer to orangutans than humans are in slow evolving genes. Among fast evolving genes, 14 showed higher identity between orangutans and gorillas than between orangutans and humans while 16 showed less (P≫0.05, Table 1). In contrast, among slow evolving genes, 27 showed higher identity between orangutans and gorillas than between orangutans and humans while 7 showed less (P=0.02), suggesting that orangutans are significantly closer to gorillas than to humans. Thus, human is the sister taxon to an orangutan-gorilla clade. The divergence time of humans and orangutans was next calculated using the fossil estimate of the gorilla split of 12 Myr ago as calibration point [18]. Assuming a constant substitution rate for the orangutan lineage, I calculated a human split of (17.57±6.9) Myr ago (Table 1).

Orangutans were also found to be closer to chimpanzees than to humans. As shown in Table 1, among fast evolving

Table 1 Orangutans are closer to gorillas or chimpanzees than to humans but are equidistant to gorillas and chimpanzees[a)]

	# Identical amino acids			Identity (%)	Divergence time (Myr)
	Or.-Hu.	Or.-Ch.	Or.-Go.	Or.-Go.	Or.-Hu.
Or.-Go.>Or.-Hu., slow evolving, 27 genes					
APOE	310	312	311/317	98	14.1
MBP1	228	228	229/235	97	14.1
KLK3	175	175	178/180	98	30.5
T2R38	298	298	299/310	96	13.1
ASIP	126	129	129/132	97	24.3
WNT7A	346	349	349/349	100	ni
FSHB	127	127	128/129	99	24
GSC	254	255	255/257	99	18.1
Myostatin	374	374	375/375	100	ni
GPR56	667	671	670/687	97	overlap
BRCA1	1098	1110	1108/1141	96	overlap
RNAseA1	149	150	151/156	96	16.7
MAOA	101	102	102/103	99	24
HNMT	112	112	113/117	96	15
SCML2	175	176	176/176	100	ni
CXCR4	346	346	347/347	100	ni
UTY	210	214	217/226	96	21.5
CFTR	1464	1465	1466/1480	99	16
Oxytocin re.	283	285	284/289	98	14.4
CXCR2	340	340	342/355	96	14
ASPM	3393	3398	3395/3447	98	12.5
CCR5	349	351	351/352	99	36.7
FUT2	330	330	331/343	96	13
Prion	248	248	249/253	98	15
TPMT	235	237	236/245	96	13.3
Globin a2	137	138	139/141	97	24.7
COX1	494	494	497/512	97	overlap
Or.-Go.<Or.-Hu., slow evolving, 7 genes					
CHRM5	290	278	286/296	96	6.0ni
MET	1382	1383	1380/1390	99	9.6ni
HTR1F	362	362	359/365	98	6.0ni
CHRM3	582	582	580/590	98	9.6
FMO 2	527	527	525/535	98	9.6
A4GALT	214	212	211/218	99	6.9ni
CORTBP2	1638	1635	1633/1663	98	10
Average					17.6±6.9
Or.-Go.>Or.-Hu., fast evolving, 14 genes					
ND2	297	299	298/346	86	
APOBEC3G	334	335	335/384	87	
COX2	214	220	219/227	94	
COX3	241	241	243/261	93	
Trim5	461	465	466/493	94	
ND6	164	164	166/174	94	
COB	339	339	342/378	90	
MCPH1	801	806	805/839	95	
MAPT	454	454	455/480	94	
NACA2	199	204	201/210	95	
SEMG2	427	ni	428/459	93	
Saitohin	119	120	121/128	94	
T2R10	234	235	236/248	95	
T2R48	257	255	258/280	92	

(*To be continued on the next page*)

(*Continued*)

	# Identical amino acids			Identity (%)	Divergence time (Myr)
	Or.-Hu.	Or.-Ch.	Or.-Go.	Or.-Go.	Or.-Hu.
Or.-Go.<Or.-Hu., fast evolving, 16 genes					
MRGX2	316	314	313/330	95	
Elafin	111	111	110/117	94	
Leptin	141	141	140/146	95	
T2R41	282	281	280/307	91	
T2R5	286	282	284/299	94	
T2R4	268	268	263/277	95	
Twist	193	190	185/203	91	
Rh50	388	387	385/409	94	
MC1R	305	305	296/317	93	
OR1D2	279	279	275/313	87	
ND5	498	496	485/585	83	
ND4	407	404	403/458	88	
ND1	277	273	274/318	87	
ATP6	188	188	181/226	80	
RNAse3	131	131	130/153	85	
T2R14	282	279	280/318	88	

a) Protein sequences from orangutans were randomly retrieved from GenBank and used to BLASTP human, chimpanzee, and gorilla protein databases at NCBI. Among the 64 informative proteins listed here, about half (30) were arbitrarily grouped as fast evolving genes based on the percentage identity between orangutans and gorillas being equal to or lower than 95%. Divergence time between orangutan and human was calculated based on the fossil split time of gorilla of 12 Myr ago. The average divergence time was calculated using slow evolving genes. Four genes from the list showing greater similarity between orangutans and chimpanzees are excluded in the calculation because they are non-informative (ni) due to 100% identity between orangutans and gorillas. To compensate for this loss of genes showing the greatest time of split between orangutans and humans, four genes from the list showing less similarity between orangutans and gorillas are also excluded, which show the smallest distance between orangutans and humans. Also, three genes with overlap ratio >0 were excluded as non-informative.

genes, 8 showed higher identity between orangutans and chimpanzees than between orangutans and humans while 10 showed less ($P \gg 0.05$). In contrast, among slow evolving genes, 17 showed higher identity between orangutans and chimpanzees while 3 showed less ($P<0.05$).

To independently verify the closer relationship between orangutans and chimpanzees, I randomly picked 733 cDNA sequences of *Pongo abelli* that were randomly generated by the German cDNA consortium. About 29.7% of these were informative (Table S13). Among fast evolving genes, 66 showed higher identity between orangutans and chimpanzees while 83 showed less ($P=0.35 \gg 0.05$). In contrast, among slow evolving genes, 53 showed higher identity between orangutans and chimpanzees while 15 showed less ($P=0.001$). Furthermore, calculations based on these slow evolving genes, assuming a 12 Myr split from orangutan for the African ape clade, gave a human split from orangutan of (17.3±5.1) Myr ago.

To verify that results from a small set of genes is representative of a much larger set of genes or even the whole genome, I analyzed all available 4330 cDNAs of *Pongo abelli* available at the GenBank. I arbitrarily divided these cDNAs into 10 groups, with every 433 cDNAs forming a group based on their numerical order of listing in the GenBank. As shown in Table S14, for fast evolving genes, 2 groups (groups 2 and 10) showed that orangutan is slightly closer to chimpanzees than to humans while 8 groups showed that orangutan is slightly closer to humans than to chimpanzees ($P>0.05$). In contrast, for slow evolving genes, all 10 groups showed that orangutan is closer to chimpanzees than to humans ($P<0.05$). None of the 10 groups individually showed that orangutan is non-equidistant to humans and chimpanzees in fast evolving genes based on the P-value cutoff of 0.05. However, for slow evolving genes, 6 groups (groups 1 and 3–7) each individually showed that orangutan is significantly closer to chimpanzees than to humans. The combined result of the 10 groups of fast evolving genes is non-significant (335 vs. 384, $P>0.05$). In contrast, the combined result of the 10 groups of slow evolving genes is extremely significant (247 vs. 80, $P<0.0001$).

To further confirm that humans are the sister taxon to a pongid clade, I determined the genetic distance to gorillas of humans and chimpanzees using a set of randomly selected gorilla proteins (Table S15). Among fast evolving proteins, 18 showed higher identity between gorillas and chimpanzees than between gorillas and humans while 16 showed less ($P \gg 0.05$). Among slow evolving genes, 27 showed higher identity between gorillas and chimpanzees while 8 showed less ($P=0.03$). The data thus show a sister grouping of gorillas and chimpanzees to the exclusion of humans.

2.3.2 Orangutans are the outgroup to a gorilla-chimpanzee clade

I next determined the relationship of the three great apes of the pongid clade using the data shown in Table 1. Among

fast evolving genes, 11 showed higher identity between orangutans and gorillas than between orangutans and chimpanzees while 18 showed less ($P \gg 0.05$). Similarly, among slow evolving genes, 12 showed higher identity between orangutans and gorillas while 14 showed less ($P \gg 0.05$). Therefore, orangutans are equidistant to gorillas and chimpanzees in both fast and slow evolving genes and are therefore the outgroup to a gorilla-chimpanzee clade.

2.3.3 Gibbons are the outgroup to a pongid-human clade

Similar analysis confirmed that the lesser ape gibbons (*Hylobates lar*) are the outgroup to a pongid-human clade (Table S16). Among fast evolving proteins, 9 showed higher identity between gibbons and orangutans than between gibbons and humans while 15 showed less ($P \gg 0.05$). Similarly, among slow evolving genes, 16 showed higher identity between gibbons and orangutans while 14 showed less ($P \gg 0.05$). Therefore, gibbons are equidistant to orangutans and humans in both fast and slow evolving genes. Gibbons are also equidistant to gorillas and humans as well as equidistant to chimpanzees and humans (data not shown).

2.3.4 Old World monkeys are the outgroup to an ape-human clade

Gibbons and humans are equidistant to the Old World monkey (OWM) *M. mulatta* in both fast and slow evolving genes (Table S17). Together with the well-established closer sequence similarity between humans and gibbons, the data suggest that monkeys are an outgroup to a clade containing gibbons and humans.

2.3.5 New World monkeys are the outgroup to an Old World monkey-human clade

Old World monkeys and humans are equidistant to New World monkeys (NWM) in both fast and slow evolving genes (Table S18). Together with the well-established closer sequence similarity between humans and OWM, the data suggest that NWM are the outgroup to a clade containing OWM and humans.

2.3.6 Simian primates are the sister taxon to a loris-tarsier clade

As shown in Table S19, among fast evolving genes, 10 showed higher identity between lorises and tarsiers than between lorises and humans while 8 showed less ($P \gg 0.05$). In contrast, among slow evolving genes, 19 showed higher identity between lorises and tarsiers than between lorises and humans while only 3 showed less ($P<0.05$), suggesting a loris-tarsier clade to the exclusion of higher primates. As an independent confirmation of this important conclusion, Table S19 also shows that loris is closer to tarsier than to the New World monkey marmoset *C. jacchus* in slow evolving genes (17 vs. 3, $P<0.05$) but not in fast evolving genes (10 vs. 5, $P>0.05$).

2.3.7 Lorises are the outgroup to a simian primate clade

Table S19 also shows that lorises are the outgroup to a simian primate clade. Among fast evolving genes, 6 showed higher identity between lorises and New World monkeys than between lorises and humans while 10 showed less ($P \gg 0.05$). Similarly, among slow evolving genes, 9 showed higher identity between lorises and New World monkeys while 11 showed less ($P \gg 0.05$). The data show that lorises are equidistant to New World monkeys and humans and are therefore the outgroup to a New World monkey-human clade given the well-established closer similarity between humans and New World monkeys than either is to lorises.

2.4 Verification of the validity and internal coherence of the primate phylogeny

I next used the newly derived human-pongid split time of 17.6 Myr, together with the well established fossil split time of 12.3 Myr between mouse and rat [26], to calculate the human-mouse divergence time. The substitution rate of the lineage leading to human was assumed to be similar to the average between human and orangutan and calculated using the human-orangutan spilt time of 17.6 Myr ($R_{human}=D/2/17.6$, where D is the distance between human and orangutan), while the substitution rate of the lineage leading to mouse was assumed to be similar to the average between mouse and rat and calculated using the mouse-rat split time of 12.3 Myr ($R_{mouse}=D/2/12.3$, where D is the distance between mouse and rat). Thus, the division time between human and mouse can be calculated as $T=D/(R_{mouse}+R_{human})$, where D is the distance between human and mouse. As shown in Table 2, a group of randomly selected slow evolving genes gave a human-mouse divergence time of 67.8 Myr, consistent with the fossil based estimation of mammalian radiation around the K-T boundary 65.5 Myr ago [22,28–31].

I then determined whether the molecular split time of 67.8 Myr between human and mouse is consistent with the fossil split time between Eutheria and Metatheria mammals [25,26]. The substitution rate of the lineage leading to Eutheria mammals was assumed to be similar to the average between human and mouse lineages during their 67.8 Myr of separation and calculated as $R_{eutheria}=D/2/67.8$, where D is the distance between human and mouse. The substitution rate of the lineage leading to Metatheria was assumed to be similar to the average between kangaroo and opossum during their 66.4 Myr of separation as determined from the fossil record [26] and calculated as $R_{metatheria}=D/2/66.4$, where D is the distance between kangaroo and opossum.

For fossil time, I assume that the real time is close to the minimum constraint time plus 10% of the minimum time, e.g., the minimum age of gorilla is 10.5 Myr and its real age is estimated as 12 Myr [18]. If such time calculation happens to be close to the maximum constraint time such as in the case of mouse-rat fossil split (minimum 11.0 vs. maxi-

Table 2 Divergence time between humans and mice[a)]

	# Identical amino acids			Divergence time of human-mus (Myr)		
	Hu-Or	Hu-Mus	Mus-rat	Hu/17.57	Mus/12.3	Hu/Mus
WNT7A	346	344	348/349	29.3	61.5	39.7
Wnt1	367	366	367/370	23.4	16.4	19.3
CAH93506	738	725	738/739	246.0	172.2	202.6
CAH90891	531	521	532/535	61.5	57.4	59.4
CAH90590	216	211	215/217	105.4	36.9	54.7
CAH93476	416	402	417/420	79.1	73.8	76.3
CAH93429	206	204	206/207	52.7	36.9	43.4
CAH93390	470	465	470/471	105.4	73.8	86.8
CAH93367	471	450	468/473	202.1	56.6	88.4
CAH93330	325	317	324/327	87.9	41	55.9
CAH93284	544	543	545/546	26.4	36.9	30.7
CAH93155	674	667	673/679	overlap		
CAH93143	322	320	323/325	29.3	30.8	30.0
CAH92769	336	332	337/338	52.7	73.8	61.5
CAH92767	1138	1136	1136/1140	35.1	12.3	18.2
CAH92738	365	358	364/366	140.6	49.2	72.9
CAH92650	224	196	224/226	263.6	184.5	217.1
CAH92595	1223	1223	1220/1230	17.6	8.6	11.6
CAH92324	906	893	904/911	63.3	31.6	42.2
CAH92088	813	800	815/819	55.6	58.4	57.0
CAH92076	513	505	512/515	87.9	41	55.9
CAH92050	336	321	335/338	149.3	69.7	95.0
CAH92747	342	338	342/343	87.9	61.5	72.4
Divergence time average (Myr)				91.0±69.9	58.4±43.0	67.8±51.5

a) Slow evolving orthologous genes were randomly selected from the German pongo cDNA project that show 99% identity between human and orangutan. Among these, some show lineage specific rate acceleration with a distance between mouse and rat that is 2-fold more than that between human and orangutan and were therefore excluded as non-neutral clock genes. Divergence time between human and mouse was calculated for each gene as shown by using human substitution rate for both lineages (Hu/17.57), mouse substitution rate for both lineages (Mus/12.3), or using human substitution rate only for the lineage leading to human and mouse substitution rate only for the lineage leading to mouse (Hu/Mus). One gene with overlap ratio >0 was excluded in the calculation.

mum 12.3 Myr), I use the maximum time. If it is close to the average of minimum and maximum, I use the average such as in the case of kangaroo and opossum (minimum 61.5 vs. maximum 71.2 Myr, average 66.4). As shown in Table 3, a group of randomly selected slow evolving genes gave a human-opossum split time of (160.4±92.9) Myr, in good agreement with the fossil record of ~160 Myr [32]. Here however the result may have some uncertainties, one due the uncertain assumption about the rate of Eutheria mammals being similar to the average between human and mouse, and the other due to the small number of available genes sampled.

3 Discussion

Past studies used average distance of all sampled genes to infer genealogy [33]. This cannot be informative because the average distance is more heavily determined/weighted by fast evolving genes that tend to show greater distances. Since previous studies made no distinction between fast and slow evolving genes, it is not unexpected that the novel results here were not found before.

3.1 Genetic non-equidistance is distinct from what is known as ‘variable molecular clock’

The variable molecular clock concept is mainly associated with two kinds of results. The first is the greater genetic distance between two sister taxa such as mouse and rat than between two other sister taxa such as human and gibbons even though the two rodents have diverged more recently based on the fossil records. The second result is related to the maximum genetic equidistance to a simpler outgroup in fast evolving genes. Some of the slight differences in distance are interpreted by the existing framework to represent significant variations in ‘mutation rate’. Thus, the variable molecular clock associated with the second result represents a kind of ‘genetic non-equidistance (to a simpler outgroup) despite equidistance in time’, which is distinct and must be differentiated from the ‘genetic non-equidistance (to a complex outgroup) despite equidistance in time’. The former is not as real as the latter and may be merely insignificant variations in maximum genetic distance (to a simpler outgroup).

Table 3 Opossum and human divergence time[a)]

	# Identical amino acids			Divergence time of Opo-Human (Myr)		
	Kan-Opo	Mus-Hu	Opo-Hu	Opo/66.4	Hu/67.8	Opo/Hu
Capza2	284	281	277/286	298.8	122.0	173.3
AAA62345	160	160	159/161	132.8	135.6	134.2
ACG50801	236	233	211/240	481.4	280.9	354.8
Mkrn1	419	406	376/428	Overlap		
G6PD	500	481	476/515	overlap		
GAPDH	220	216	217/228	91.3	62.2	74.0
ACM88712	174	176	171/181	94.9	135.6	111.6
PR	172	170	157/180	190.9	155.9	171.7
Pgk1	390	407	383/416	overlap		
UBE1y1	141	149	144/152	overlap		
Cav1	165	169	160/178	overlap		
PRDX1	182	189	180/198	overlap		
ABW82472=prdx1	182	189	183/198	overlap		
Cox1	493	466	459/512	overlap		
CytoC	101	96	95/105	166	75.3	103.6
Divergence time average (Myr)				208.0±139.7	138.2±71.5	160.4±92.9

a) Slow evolving orthologous genes with greater than 90% identity between kangaroo (*Macropus eugenii*) and opossum (*Monodelphis domestica*) and between human and mouse were randomly selected from the NCBI database. All informative genes available from the database are included in the table. Genes showing lineage specific mutation rate acceleration were non-informative and excluded. Divergence time between human and opossum was calculated for each gene as shown by using opossum substitution rate for both lineages (Opo/66.4), human substitution rate for both lineages (Hu/67.8), or using opossum substitution rate only for the lineage leading to opossum and human substitution rate only for the lineage leading to human (Opo/Hu). Genes with overlap ratio >0 were excluded from the calculation.

More importantly, it also must be differentiated from the 'real' non-equidistance to a simpler taxon associated with non-equidistance in time in slow evolving genes.

While humans and chimpanzees are approximately equidistant to orangutans as measured by fast evolving intron and intergenic regions, humans can be shown to be slightly closer to orangutans [34]. Since the genetic diversity of chimpanzees is higher than that of humans, chimpanzees contribute slightly more than humans to the maximum distance with orangutans. Because the difference is extremely small, it requires large amount of fast evolving sequences to be seen. In contrast, the real non-equidistance of humans and chimpanzees to orangutans can be easily shown using only ~20 slow evolving proteins.

3.2 The meaning of 'most recent common ancestor'

Based on the fossil record, there exist two kinds of diversification from an ancestor. One is slow and gradual and the other is fast and explosive. From fish to amphibian is a slow process. The oldest fish fossil is ~530 Myr old while the oldest amphibian fossil is ~340 Myr old. Here the most recent common ancestor (MRCA) of fish and amphibian is an individual fish from ~340 Myr ago. This MRCA would account in theory for all extant amphibians but only a tiny fraction of all extant fishes. In contrast, when diversification proceeds via radiation or explosion, the MRCA of two extant species may account for all living individuals of these two species and may not look like either species. For speciation via radiation, one may not be able to identify the MRCA fossil since it may not look like any living species. However, for gradual speciation, one extant species would have existed longer than another. Here, the oldest fossil for the older lineage would not be informative to divergence time but only the oldest fossil of the younger lineage will. A fossil species can be either sister or ancestor to a living species. While the positive identification of a fossil as a sister of a living species by cladistic analysis also implies a possibility for it to be an ancestor, a failure to do so cannot exclude it as an ancestor.

The presently popular notion of MRCA is needed in order to make sense of the molecular data in terms of the molecular clock/Neutral theory. If the MRCA of sister taxa A and B was a member of B, with B being the older and less derived/complex lineage of the two sisters, and only appeared many years after the existence of B and produced a fraction of extant B represented by B1 (Figure 5), then it is possible for some extant members of B (B2 in Figure 5) to have a separation time with B1 that is longer than that between A and B1. Then, the largest genetic distance within B in most sequences would be greater than the minimum distance between A and B, according to the molecular clock/Neutral theory (Figure 5). In other words, the genetic diversity of B would be greater than that of the clade containing A and B1, since B has existed longer and accumulated more mutations. However, the fact is that the largest genetic distance within a taxon in most sequences (keep in mind that most sequences are fast evolving) is mostly not greater than the distance between the taxon and its sister taxon.

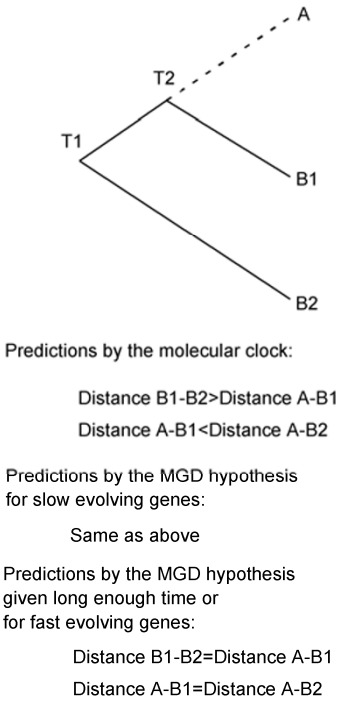


Figure 5 The concept of most recent common ancestor. B1 and B2 are extant individuals of taxon B and shared a common ancestor at time T1. Taxon A is a more derived sister taxon of B and shared a common ancestor with a fraction of B (B1) at time T2. The difference in time between T1 and T2 can be from zero to any size. B-like lineage is represented by solid line while A-like lineage by dashed line. Between T1 and T2, the line leading to A is still part of the B-like lineage. The predictions by the molecular clock/neutral theory and the MGD hypothesis are shown.

Therefore, the traditional notion requires that the direct ancestor of A and the direct ancestor of B, both being members of the B lineage, were either the same individual or had not lived many years apart. The MGD hypothesis however does not require that B cannot appear many years prior to the MRCA of A and B1 (Figure 5). For most sequences, MGD would have been reached within B, which could never be greater than the minimum genetic distance between A and B since the distance between A and B should be similar to the MGD of B. The new MRCA concept for gradual diversifications suggests that while the direct ancestor of A was an individual (or pair) from the B-like lineage, the ancestors of B were many individuals from the B-like lineage living at different times (Figure 5). The MRCA of A and B1 marked the first appearance of A or A-like at time T2 in Figure 5 but not the first appearance of B or B-like lineage, which first appeared at time T1 in Figure 5. It accounts for all extant individuals of A but only a fraction of all extant individuals of B if some of its descendants had remained as B. This new MRCA concept is consistent with the trend in gradual diversification that one of the sister taxa is often more similar than the other to the ancestor lineage in morphology. Gorillas are the sister taxon of chimpanzees and are more similar to orangutans [35]. Similarly, the MRCA of human and pongids should be a member of an orangutan-like lineage. Thus, humans should share similarity with orangutans, which in fact is the case [4,6,36].

3.3 The slow clock method and the new primate phylogeny

The slow clock method here has two premises. The first is that sequence similarity sometimes (not always) reflects genealogical relationship, which is an easily proven fact well described by the Neutral theory. Using only slow evolving proteins that show zero overlapped or coincident substitutions insures that the relationship between time and distance is truly linear. The second is the approximate constancy of neutral substitution rate in protein or DNA sequence within the neutral diversity range for any single lineage over its evolutionary life time.

The new primate phylogeny is shown in Figure 6. The results here found no evidence of a gorilla-chimpanzee-human clade with orangutan as the outgroup, a chimpanzee-human clade with gorilla as the outgroup, or a tarsier-simian primate clade with lorises as the outgroup, all controversial clades claimed by the old approach but either contradicted or unresolvable by the fossil records. In contrast, the same method positively identified an orangutan-gorilla-chimpanzee clade with human as the outgroup, a gorilla-chimpanzee clade with orangutan as the outgroup, and a loris-tarsier clade with simian primates as the outgroup, all consistent with paleontological findings or tradi-

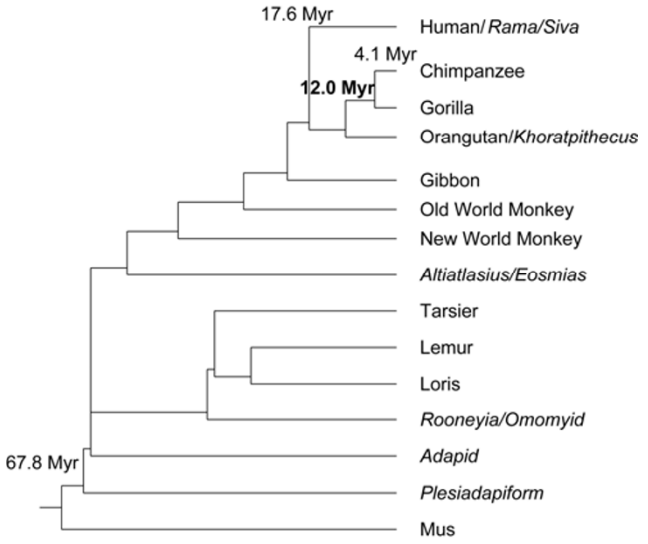


Figure 6 A phylogeny of primates. The relationships of selected major primates are shown, based on results of this study. The shorter vertical distance between an MRCA node and one of the sister taxa indicates that a past member of that taxon was the ancestor of the MRCA. For example, a past member of the gorilla lineage was the ancestor of the MRCA shared by all extant chimpanzees and a fraction of extant gorillas. Divergence times calculated by the slow clock method are indicated and that in bold represents fossil times used as calibration for the slow clock.

tional views of paleontologists before the molecular clock era. The positive identification of well-established clades serves as a powerful positive control for the method here and the conclusions on the controversial clades. In contrast, traditional methods often misidentify well-established relationships, and its rationale for human-chimpanzee grouping would also misplace humans with octopus to the exclusion of cockles, or with birds to the exclusion of snakes. By deducing, from ~17.6 Myr divergence time between human and pongid, the molecular time for mammal radiation and for Eutheria-Metatheria split that are consistent with the fossil records, the results indicate an internal coherence of the primate phylogeny.

The new molecular results here strongly support the original view of paleoanthropologists on *Ramapithecus* [1–3]. As a very early human ancestor, it should resemble orangutans and it does. The results also easily accommodate the 7 Myr old hominid *Sahelanthropus* [17,37] and the 5.6–4.4 Myr old *Ardipithecus* [38].

There are a number of different fossil apes around 17–15 Myr ago in Africa that can be divided roughly into two major groups according to some authors [39,40]. Group one consists of *Turkanapithecus* and *Kenyapithecus*, and the other group of *Afropithecus*, *Equatorius*, and *Nacholapithecus*. Group one has no suspension adaptation in locomotion and may have migrated to Eurasia around 15–14 Myr ago and given rise to one of the two types of *Griphopithecus* and *Sivapithecus* who later may have moved back to Africa around 8–10 Myr ago due to climate change to a temperate one in Eurasia. Group two also may have moved to Eurasia around 15–14 Myr ago and given rise to the other type *Griphopithecus* (*G. alpani*) and *Dryopithecus*. Change to temperate climate in late Miocene in Eurasia may have caused some *Dryopithecus* to move back to Africa around 12 Myr ago leading to African apes and some (*D. laietanus*) to tropical South East Asia leading to *Khoratpithecus* and orangutans. The African ape ancestors may be more sensitive to temperate climates and disappearance of forests than human ancestors and thus moved back to Africa earlier, at the beginning period of climate cooling.

The two groups of fossil apes at 14–10 Myr ago are more distinctly different than their earlier African ancestors [40]. According to Stringer and Andrews: "Group one has robust jaws, enlarged molar teeth with thick enamel, and some buttressing of the face to accommodate chewing stresses caused by the large teeth and a hard fruit diet. They lived in seasonal woodland to open forest environments and were adapted to some extent to ground living [40]." They were likely the ancestors of humans and later developed bipedalism and gave rise to *Ardipithecus*. To some authors, walking on two legs may arise more likely from a terrestrial form of locomotion on all fours rather than arboreal climbing and suspension [41,42]. "The other group inhabited wetter, less seasonal forests and lived in trees employing a form of locomotion that involves some degree of suspension from overhead branches [40]." They were obviously the best candidates for the ancestors of pongids.

Chimpanzees had lived side by side with humans in the past in areas suitable for fossil formations [43]. The emergence of chimpanzees from a gorilla-like lineage was here calculated to be 4.1 Myr ago (Table S18). The only known ancient fossil of chimpanzees has an age of 0.5 Myr [43]. The much more recent emergence of chimpanzees easily explains the extreme rarity of chimpanzee fossils relative to that of humans (or even to gorillas).

The main seeming inconsistency with this story is the intermediate thin enamel of *Dryopithecus* being unlike the intermediate thickness in orangutans and in the oldest fossil gorilla *Chororapithecus*. But enamel thickness can vary a great deal within a species [44]. Besides, the enamel thickness of orangutan is really an intermediate between human and African apes [44], and its enamel deposition rate is slow like African apes rather than fast like *Sivapithecus* and humans [36].

Previous molecular studies including the analysis of Alu insertions show that tarsiers are closer in sequence to simian primates than lemurs/lorises are. But that is likely due to convergent evolution, because tarsiers show more features of higher epigenetic complexity than other prosimians, including long gestation time and brain at birth largest among mammals relative to body size [20].

There are ample molecular data supporting the pongid clade. First, chimpanzee is closer to orangutan or gorilla than human is in gene expression patterns [45–47]. Second, the chromosome-banding pattern of humans is more similar to orangutan than to chimpanzee or gorilla [48]. Finally, human-specific segmented duplications show lower copy number polymorphisms in humans than chimpanzee-specific segmented duplications do in chimpanzees [49]. Similarly, those duplications shared among human, chimpanzees, and orangutans, or those shared among human, chimpanzees, orangutans, and monkeys are also less polymorphic in humans than in chimpanzees, indicating that duplications that are shared because of common ancestry are less polymorphic in humans than in chimpanzees. In contrast, the duplications shared between human and chimpanzees are equally polymorphic in humans and chimpanzees. This unusual result contradicts the sister grouping of humans and chimpanzees, because either the MGD or the alternative bottleneck hypothesis would predict lower polymorphism in humans if these duplications are shared because of common ancestry. However, it is fully consistent with the interpretation that the shared duplications between humans and chimpanzees are not due to common ancestry but are due to common selection of independent duplications. Common selection leading to shared sequences is well established [50–52]. The MGD hypothesis interprets

many of the shared sequences between humans and chimpanzees as a result of common selection rather than common ancestry. Similar selection pressures lead to similar levels of polymorphism. This result is thus one of the best that cannot be reconciled with the sister grouping of humans and chimpanzees but strongly supports the sister grouping of humans and pongids.

This work was supported by the State Key Laboratory of Medical Genetics, a FuRong Scholarship, the National Natural Science Foundation of China (Grant No. 81171880) and the National Basic Research Program of China (Grant No. 2011CB51001). I thank Zhou WenYun and Zeng Ceng for technical assistance and John Grehan for reading the manuscript and for valuable discussions.

1 Simons E L. The phyletic position of Ramapithecus. Postilla, 1961, 57: 1–9
2 Simons E L, Pilbeam D R. Preliminary revision of the Dryopithecinae (Pongidae, Anthropoidea). Folia Primatol (Basel), 1965, 3: 81–152
3 Pilbeam D. The earliest hominids. Nature, 1968, 219: 1335-1338
4 Schwartz J H. The evolutionary relationships of man and orang-utans. Nature, 1984, 308: 501–505
5 Lewin R. Human Evolution. 5th ed. Malden: Blackwell Publishing Ltd., 2005
6 Schwartz J H. The Red Ape, Orangutans and Human Origins. Cambridge: Westview Press, 2005
7 Goodman M. Immunochemistry of the primates and primate evolution. Ann N Y Acad Sci, 1962, 102: 219–234
8 Sarich V M, Wilson A C. Immunological time scale for hominid evolution. Science, 1967, 158: 1200–1203
9 Wilson A C, Sarich V M. A molecular time scale for human evolution. Proc Natl Acad Sci USA, 1969, 63: 1088–1093
10 Margoliash E. Primary structure and evolution of cytochrome c. Proc Natl Acad Sci USA, 1963, 50: 672v679
11 Huang S. The overlap feature of the genetic equidistance result, a fundamental biological phenomenon overlooked for nearly half of a century. Biol Theory, 2010, 5: 40–52
12 Kimura M. Evolutionary rate at the molecular level. Nature, 1968, 217: 624–626
13 Huang S. Histone methylation and the initiation of cancer. In: Tollefsbol T, ed. Cancer Epigenetics., New York: CRC Press, 2008
14 Huang S. Inverse relationship between genetic diversity and epigenetic complexity. Preprint available at Nature Precedings 2009. http://dx.doi.org/10.1038/npre.2009.1751.2
15 Copley R R, Schultz J, Ponting C P, *et al*. Protein families in multicellular organisms. Curr Opin Struct Biol, 1999, 9: 408–415
16 Huang S. The genetic equidistance result of molecular evolution is independent of mutation rates. J Comp Sci Syst Biol, 2008, 1: 92-102
17 Brunet M, Guy F, Pilbeam D, *et al*. A new hominid from the Upper Miocene of Chad, Central Africa. Nature, 2002, 418: 145–151
18 Suwa G, Kono R T, Katoh S, *et al*. A new species of great ape from the late Miocene epoch in Ethiopia. Nature, 2007, 448: 921–924
19 Shoshani J, Groves C P, Simons E L, *et al*. Primate phylogeny: morphological vs. molecular results. Mol Phylogenet Evol, 1996, 5: 102–154
20 Schwartz J H. How close are the similarities between Tarsius and other primates? In: Wright P C, Simons E L, Gursky S, eds. Tarsiers: Past, Present and Future. Piscataway: Rutgers University Press, 2003
21 Bininda-Emonds O R, Cardillo M, Jones K E, *et al*. The delayed rise of present-day mammals. Nature, 2007, 446: 507–512
22 Wible J R, Rougier G W, Novacek M J, *et al*. Cretaceous eutherians and Laurasian origin for placental mammals near the K/T boundary. Nature, 2007, 447: 1003–1006
23 Flynn J J, Parrish J M, Rakotosamimanana B, *et al*. A new Middle Jurassic mammals from Madagascar. Nature, 1999, 401: 57–60
24 Kumar S, Hedges S B. A molecular timescale for vertebrate evolution. Nature, 1998, 392: 917–920
25 Luo Z X, Ji Q, Wible J R, *et al*. An Early Cretaceous tribosphenic mammal and metatherian evolution. Science, 2003, 302: 1934–1940
26 Benton M J, Donoghue P C. Paleontological evidence to date the tree of life. Mol Biol Evol, 2007, 24: 26–53
27 Chatterjee H J, Ho S Y W, Barnes I, *et al*. Estimating the phylogeny and divergence times of primates using a supermatrix approach. BMC Evol Biol, 2009, 9: 259
28 Bajpai S, Kay R F, Williams B A, *et al*. The oldest Asian record of Anthropoidea. Proc Natl Acad Sci USA, 2008, 105: 11093–11098
29 Sige B, Jaeger J J, Sudre J, *et al*. Altiatlasius koulchii n. gen. et sp., primate omomyidé du Paléocène supérieur du Maroc, et les origines des euprimates. Palaeontographica Abt A, 1990, 214: 31–56
30 Beard C. The Hunt for the Dawn Monkey. Berkeley: University of California Press, 2004
31 Kay R F, Ross C, Williams B A. Anthropoid origins. Science, 1997, 275: 797–804
32 Luo Z X, Yuan C X, Meng Q J, *et al*. A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature, 2011, 476: 442–445
33 Wildman D E, Uddin M, Liu G, *et al*. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo. Proc Natl Acad Sci USA, 2003, 100: 7181–7188
34 Elango N, Thomas J W, Yi S V. Variable molecular clocks in hominoids. Proc Natl Acad Sci USA, 2006, 103: 1370–1375
35 Collard M, Wood B. How reliable are human phylogenetic hypotheses? Proc Natl Acad Sci USA, 2000, 97: 5003–5006
36 Grehan J R. Mona Lisa smile: the morphological enigma of human and great ape evolution. Anat Rec B New Anat, 2006, 289: 139–157
37 Lebatard A E, Bourles D L, Duringer P, *et al*. Cosmogenic nuclide dating of *Sahelanthropus tchadensis* and *Australopithecus bahrelghazali*: Mio-Pliocene hominids from Chad. Proc Natl Acad Sci USA, 2008, 105: 3226–3231
38 White T D, Asfaw B, Beyene Y, *et al*. Ardipithecus ramidus and the paleobiology of early hominids. Science, 2009, 326: 75–86
39 Cela-Conde C J, Ayala F J. Human Evolution: Trails from the Past. Oxford: Oxford University Press, 2007
40 Stringer C, Andrews P. The Completer World of Human Evolution. New York: Thames and Hudson, 2005
41 Schwartz J H. The origins of human bipedalism. Science, 2007: 1065
42 Thorpe S K, Holder R L, Crompton R H. Origin of human bipedalism as an adaptation for locomotion on flexible branches. Science, 2007, 316: 1328–1331
43 McBrearty S, Jablonski N G. First fossil chimpanzee. Nature, 2005, 437: 105–108
44 Smith T M, Martin L B, Leakey M G. Enamel thickness, microstructure and development in Afropithecus turkanensis. J Hum Evol, 2003, 44: 283–306
45 Enard W, Khaitovich P, Klose J, *et al*. Intra- and interspecific variation in primate gene expression patterns. Science, 2002, 296: 340–343
46 Uddin M, Wildman D E, Liu G, *et al*. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. Proc Natl Acad Sci USA, 2004, 101: 2957–2962
47 Karaman M W, Houck M L, Chemnick L G, *et al*. Comparative analysis of gene-expression patterns in human and African great ape cultured fibroblasts. Genome Res, 2003, 13: 1619–1630
48 Yunis J J, Prakash O. The origin of man: a chromosomal pictorial legacy. Science, 1982, 215: 1525–1530
49 Marques-Bonet T, Kidd J M, Ventura M, *et al*. A burst of segmental duplications in the genome of the African great ape ancestor. Nature, 2009, 457: 877–881
50 Bull J J, Badgett M R, Wichman H A, *et al*. Exceptional convergent evolution in a virus. Genetics, 1997, 147: 1497–1507

51 Bollback J P, Huelsenbeck J P. Parallel genetic evolution within and between bacteriophage species of varying degrees of divergence. Genetics, 2009, 181: 225–234

52 Castoe T A, de Koning A P, Kim H M, *et al*. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci USA, 2009, 106: 8986–8991

Biographical Sketch

Dr. Huang Shi was a professor at the State Key Laboratory of Medical Genetics at Central South University in Changsha, China. He grew up in the army compound of the Chinese Academy of Military Medicine in Beijing where his father was a professor. His childhood interest was however not medicine but sports and later fine arts. An unsuccessful effort in the entrance examination of the Chinese Central Academy of Fine Arts in 1978 changed his interest to science. He entered Fudan University in 1979 and graduated in 1983 with a bachelor's degree in genetics. He was a CUSBEA fellow of Class III (1984) and obtained his Ph.D. in biochemistry at the University of California at Davis in 1988. After finishing a postdoctoral training at the University of California at San Diego, he was appointed in 1992 assistant professor at the Sanford-Burnham Institute and promoted to associate professor in 1998. He was appointed professor at Central South University in 2009. The early training in art helped shape his taste in aesthetics and interest in science only as a creative endeavor. His laboratory discovered the RIZ or PRDM family of histone methyltransferases and proposed an epigenetic pathway of carcinogenesis by diet rich in meat and low in vegetables. Since 2003, he initiated study of the relationship between genetics and epigenetics and its role in the evolution of biological complexity. He proposed the maximum genetic diversity hypothesis and has been using it to rewrite evolution and population genetics as well as to solve genetic puzzles of complex traits/diseases about which the existing paradigm is clueless. He was one of the 1993 class of Pew Scholars in the Biomedical Sciences.

Supporting Information

Primate phylogeny: molecular evidence for a pongid clade excluding humans and a prosimian clade containing tarsiers

1 Genetic non-equidistance to a more complex outgroup despite equidistance in time

Table S1 Human relationship with mollusks. The percentage identities in protein sequence between species (*Octopus vulgaris*, *Acanthocardia tuberculatum*, and *Homo sapiens*) are shown for 10 mitochondrial proteins

Table S2 Human relationship with brachiopods. The percentage identities in protein sequence between species (*Terebratulina retusa*, *Lingula anatina*, and *Homo sapiens*) are shown for 10 mitochondrial proteins

Table S3 The reptile clade (including birds): human is closer to birds than to snakes. The percentage identities in protein sequence between species (birds, snakes, and humans) are shown for 10 mitochondrial proteins and 13 randomly selected proteins encoded by the nuclear genome. The number was from BLASTP analysis of bird or snake database from Genbank and represent the highest identity. The mitochondrial proteins show that snakes are more distant to humans than birds are (P<0.05). A random sampling of 13 nuclear genes also showed the same result (P<0.05)

Table S4 The amphibian group. The percentage identities in protein sequence between species (*Xenopus laevis*, *Limnonectes fujianensis,* and *Homo sapiens*) are shown for 12 randomly selected proteins. The number was from BLASTP analysis of GenBank. The data show that *Xenopus laevis* is closer to humans than *Limnonectes fujianensis* is, but more proteins need to be sampled to confirm the significance of this trend (P=0.06). *Limnonectes fujianensis* is closer to *Xenopus laevis* than to humans (P=0.01), consistent with a closer phylogenetic relationship between the two frogs

Table S5 The teleost fish group: human is closer to the loach than to the three spined frogfish. The percentage identities in protein sequence between species (*Vaillantella maassi*, *Batrachomoeus trispinosus*, and *Homo sapiens*) are shown for 13 mitochondrial proteins. The mitochondrial proteins show that the loach *Vaillantella maassi* is significantly closer to humans than the three spined frogfish *Batrachomoeus trispinosus* is (P=0.005). The data suggest that some teleost fishes are closer to humans than others, presumably due to higher epigenetic complexity. Future work is needed to determine if the loach is indeed more complex than the frogfish. Also, the frogfish is closer to the loach than to humans (P=0.03), consistent with a closer phylogenetic relationship between the two fishes

Table S6 The echinoderm phylum. The percentage identities in protein sequence between species (*Strongylocentrotus purpuratus*, *Ophiura lutkeni*, and *Homo sapiens*) are shown for 11 mitochondrial proteins. Using COX1 and COB proteins of humans as query, the sea urchin (*Strongylocentrotus purpuratus*) was identified as among the closest to humans, while the starfish (*Ophiura lutkeni*) was found among the most distant. A sampling of 11 proteins shows that sea urchin is slightly closer to humans than the starfish is (P=0.19). Future work with more proteins will be needed to determine if this trend is significant. The starfish is slightly closer to sea urchins than to humans (P=0.07), consistent with a clade containing the starfish and sea urchins

Table S7 The arthropod phylum: human is closer to the dragonfly than to the louse. The percentage identities in protein sequence between species (*Orthetrum triangulare melania*, *Campanulotes bidentatus compar*, and *Homo sapiens*) are shown for 10 mitochondrial proteins. The wingless louse (*Campanulotes bidentatus compar*) was identified as among the most distant to humans as measured by a randomly chosen protein COX1. The dragonfly (*Orthetrum triangulare melania*) was identified as among the closest to humans among arthropods. The distance of these two species to humans was next determined using ten mitochondrial proteins. Humans are significantly closer to the dragonfly than to the louse (P<0.05). This suggests that the dragonfly is more complex than the wingless louse, which is consistent with fact that the former can fly. The louse is not significantly closer to dragonfly than to human (P=0.35), suggesting that the distance between the two insects is close to the maximum

Table S8 The nematode phylum. The percentage identities in protein sequence between species (*Cooperia oncophora*, *Brugia malayi*, and *Homo sapiens*) are shown for 10 randomly selected proteins. Using COX1 and COB proteins of humans as query, *Cooperia oncophora* was identified as among the closest to humans, while *Brugia malayi* was found among the most distant. A sampling of 11 proteins showed that there is a trend (P=0.06) for a closer relationship between *Cooperia oncophora* and human. *Brugia malayi* is not significantly closer to *Cooperia oncophora* than to humans, suggesting that the distance between the two nematodes is close to the maximum

Table S9 The porifera phylum: human is closer to the chicken liver sponge than to *H. lachne.* The percentage identities in protein sequence between species (*Chondrilla aff. nucula*, *Hippospongia lachne*, and *Homo sapiens)* are shown for 10 mitochondrial proteins. Using COX1 and COB proteins of humans as query, the chicken liver sponge (*Chondrilla aff. nucula*) was identified as among the closest to humans, while *Hippospongia lachne* was found among the most distant. A sampling of 10 proteins showed that humans are significantly closer to *Chondrilla aff. nucula* than to *Hippospongia lachne* (P<0.05). However, *Hippospongia lachne* is not the sister taxon of a human-*Chondrilla* clade since it is closer to *Chondrilla aff. nucula* than to humans (P<0.05)

Table S10 The fungi kingdom: human is closer to the corn smut than to yeast. The percentage identities in protein sequence between species (*Ustilago maydis*, *Candida zemplinina or Candida*, and *Homo sapiens*) are shown for 20 random selected proteins. Using COX1 and COB of humans as query, the smut fungus *Ustilago maydis* was identified among the closest to humans, while the yeast *Candida zemplinina* was among the most distant to humans. A sampling of five proteins (few *C. zemplinina* protein sequences are known) showed that the smut fungus is closer to humans than the yeast. To confirm that the smut fungus is indeed closer to humans than the Candida genus, 15 more proteins were randomly sampled. Among different Candida species, the one showing the highest identity with human is shown in the Table. The smut is closer to humans than Candida is in 19 of 20 proteins (P = 0.003). The data suggest that the smut has higher epigenetic complexity than the yeast, consistent with the status of this fungus as 'Higher Fungi'. However, Candida is not an outgroup to a human-smut clade since it is closer to smut than to humans (P = 0.04)

Table S11 The protist alveolates superphylum. The percentage identities in protein sequence between species (*Plasmodium falciparum*, *Tetrahymena thermophila*, and *Homo sapiens*) are shown for 11 random selected proteins. Using COX1 of humans as query, the malaria parasite *Plasmodium* (phylum Apicomplexa) was identified among the closest to humans, while *Tetrahymena* (phylum Ciliophora) was among the most distant. However, a sampling of 11 proteins showed that, relative to *Tetrahymena, Plasmodium* is closer to humans in 5 proteins but more distant in 6 proteins. Thus, the two species are equidistant to humans. Coincidence and common selection may account for the large differences in identity to humans between the two species in some proteins such as COX1, COB, and GPDH. The two protists are also no closer than either is to humans, suggesting that the separation time for the two protists has been long enough for their genetic distance to reach the maximum cap

2 Difference between slow and fast evolving genes in phylogeny inference

Table S12 Difference between slow and fast evolving genes in phylogeny inference. The percentage identities between zebrafish (*D. rerio*) and pufferfish (*T. nigroviridis*), human (*H. sapiens*), or mouse (*M. musculus*) are shown for a number of lysine methyltransferases (KMTs) and ribosome proteins. Genes are considered as having reached maximum distance in fishes if the identity between the two fishes is equal to or slightly smaller than that between fish and mammal

3 Pongo abelli is closer to Pan troglodytes than to Homo sapiens

Table S13 Divergence time between homo and *Pongo abelli* based on sequences from *Pongo* abelli, *Pan troglodytes*, and *Homo sapiens*. Of 733 randomly selected cDNA sequences from *P. abelli* (NCBI accession number, CAI29673 to CAI29581, CAH93520 to CAH93492, CAH92004 to 91825, CAH91005 to CAH90750, and CAH90602 to CAH90424), 218 sequences are informative and listed here. 68 have greater than 98% identity between *P. abelli* and *P. troglodytes* and are considered as slow evolving proteins, while the other 149 proteins have identities between *P. abelli* and *P. troglodytes* that are equal to or smaller than 98% and are considered fast evolving. Among fast evolving genes, 66 showed higher identity between orangutans and chimpanzees while 83 showed less (P = 0.35 ≫ 0.05). In contrast, among slow evolving genes, 53 showed higher identity between orangutans and chimpanzees while 15 showed less (P < 0.001)

Table S14 *Pongo abelli* is closer to *Pan troglodytes* than to *Homo sapiens*. Of 4330 random cDNA sequences of *P. abelli* available from Genbank, every 433 sequences based on their numerical order of appearance on the NCBI webpage were selected to form an experimental group. Orthologous genes with greater than 98% identity between *P. abelli* and *P. troglodytes* were considered as slow evolving proteins, while genes with identities between *P. abelli* and

P. troglodytes that are equal to or smaller than 98% are considered fast evolving. The meaning of C-O > H-O: the percentage identity between chimpanzees (C) and orangutans (O) is greater than between humans (H) and orangutans. Numbers in parenthesis indicate *P*-alues from Fisher's exact test (2 tailed)

4 Gorillas are closer to chimpanzees than to huma

Table S15 Gorillas are closer to chimpanzees than to humans. Of the 69 informative gorilla proteins listed here, 35 have greater than 97% identity between gorillas (Go) and chimpanzees (Chimp) and are considered as slow evolving proteins, while the other 34 proteins have identities between gorillas and chimpanzees that are equal to or smaller than 97% and are considered fast evolving. Among fast evolving proteins, 18 showed higher identity between gorillas and chimpanzees than between gorillas and humans while 16 showed less ($P \gg 0.05$). In contrast, among slow evolving genes, 27 showed higher identity between gorillas and chimpanzees while 8 showed less ($P = 0.03$)

5 Gibbons are the ougroup to a pongid-hominid clade

Table S16 Gibbons are equidistant to orangutans and humans. Of the 53 informative gibbon (*Hylobates lar*) proteins shown here, 19 have greater than 95% identity between gibbons (Gi) and orangutans (Orang) and are considered slow evolving, while the other 34 proteins have identities between gibbons and orangutans that are equal to or smaller than 95% and are considered fast evolving. Among fast evolving proteins, 13 showed higher identity between gibbons and orangutans than between gibbons and humans while 21 showed less ($P \gg 0.05$). Similarly, among slow evolving genes, 12 showed higher identity between gibbons and orangutans than between gibbons and humans while 7 showed less ($P \gg 0.05$). The data show that gibbons are equidistant to orangutans and humans in both slow and fast evolving genes

6 Old World monkeys are the outgroup to an ape-human clade

Table S17 Old World monkeys are equidistant to gibbons and humans. Of the 34 informative Old World monkeys (macaque) proteins shown here, 18 have greater than 92% identity between macaque (Ma) and gibbons (Gi) and are considered slow evolving, while the other 16 proteins have identities between macaque and gibbons that are equal to or smaller than 92% and are considered fast evolving. Among fast evolving proteins, 8 showed higher identity between macaques and gibbons than between macaques and humans while 8 showed less ($P \gg 0.05$). Similarly, among slow evolving genes, 7 showed higher identity between macaques and gibbons than between macaques and humans while 11 showed less ($P \gg 0.05$). The data show that macaques are equidistant to gibbons and humans in both slow and fast evolving genes

7 New World monkeys are the outgroup to an Old World monkey-ape-human clade

Table S18 New World monkeys are equidistant to Old World monkeys and humans. Of the 39 informative New World monkeys (Saguinus) proteins shown here, 17 have greater than 90% identity between Saguinus (Sa) and macaque (Ma) and are considered slow evolving, while the other 22 proteins have identities between Saguinus and macaque that are equal to or smaller than 90% and are considered fast evolving. Among fast evolving proteins, 9 showed higher identity between Saguinus and macaques than between Saguinus and humans while 13 showed less ($P \gg 0.05$). Similarly, among slow evolving genes, 8 showed higher identity between Saguinus and macaques than between Saguinus and humans while 9 showed less ($P \gg 0.05$). The data show that New World monkeys are equidistant to macaques and humans in both slow and fast evolving genes

8 Simian primates are the sister taxon to a loris-tarsier clade

Table S19 Lorises are closer to tarsiers than to humans but are equidistant to New World monkeys and humans. Most of the protein sequences of lorises available at the Genbank were selected for comparison with humans, tarsiers, and New World monkeys (NWM). Of the 40 informative proteins as shown here, 22 have greater than 85% identity between lorises and tarsiers and are considered slow evolving, while the other 18 proteins have identities between lorises and tarsiers that are equal to or smaller than 84% and are considered fast evolving

9 Calculation of the divergence time between chimpanzees and gorillas

Table S20 The divergence time between chimpanzees and gorillas. The 27 slow evolving genes as listed in Table 4 were used to calculate the divergence time between chimpanzees and gorillas. This calculation assumes that the mutation rates in these genes are similar in gorillas and orangutans, which is highly likely given the close relationship between the two apes. Calculation based on the gorilla fossil split time of 12 Myr ago was performed using the formula: Divergence time of chimpanzees and gorillas = 12× the Poisson correction distance between gorillas and chimpanzees divided by the Poisson correction distance between gorilla and orangutan. ni: most of the non-informative genes show 100% identity either between chimpanzees and gorillas or between gorillas and orangutans. Two genes (KLK3 and CCR5) shows more identity between gorilla and orangutans than between chimpanzees and gorillas and has likely reached cap of diversity and is therefore non-informative

The supporting information is available online at life.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.